An introduction to the `rfishbase` Package

Carl Boettiger[a,*], Peter Wainwright[a]

[a] *Center for Population Biology, University of California, Davis, United States*

Abstract

We introduce a package that provides programmatic access to the FishBase repository~[2]. This package allows us to interact with data on over 30,000 fish species in the rich statistical computing environment, `R`, and can also allow the user to integration with other research software, such as the phylogenetics package~`ape`.

Keywords: R, vignette, fishbase

1. Introduction

Describe the fishbase database ... extent, information avialable, etc Briefly describe R
Stuff about Machine access to data, role of large scale data in ecology [4], [3], [7]
Froese and Pauly. [2]

```
require(rfishbase)
require(ggplot2)
```

2. Examples

The `rfishbase` package works by creating a cached copy of all data on fishbase currently available in XML format. Caching increases the speed of queries and places minimal demands on the fishbase server, which in it's present form is not built to support direct access to application programming interfaces (APIs). The cached copy can be loaded in to R using the command:

```
data(fishbase)
```

To get the most recent copy of fishbase, update the cache instead. The update may take up to 24 hours. This copy is stored in the working directory with the current date and can be loaded when finished.

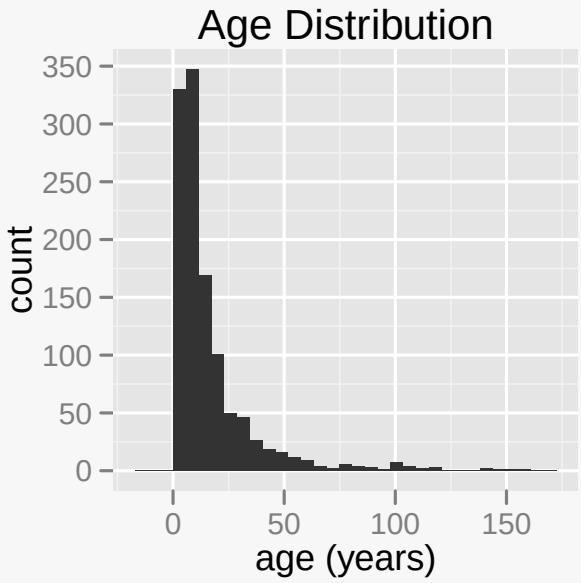
```
updateCache()
loadCache("2011-10-12fishdata.Rdat")
```

Loading the database creates an object called fish.data, with one entry per fish species for which data was successfully found, for a total of

```
length(fish.data)

[1] 30622
```

```
yr <- getSize(fish.data, "age")
nfish <- length(fish.data)
qplot(yr, main = "Age Distribution", xlab = "age (years)")
```



2.1. Simple queries

The examples we give here are meant to be illustrative of the kinds of queries that are possible, and also provide a simple introduction to assist the reader in using the software itself.

Typical use of the package constructs queries to identify those species matching certain criteria, which can easily be combined with other queries to quickly answer potentially questions that would otherwise be tedious to evaluate. For instance, we can ask "are there more labrids or goby species of reef fish?" using the following queries:

Get all species in fishbase from the families "Labridae" (wrasses) or "Scaridae" (parrotfishes), which are both labrids:

```
labrid <- familySearch("(Labridae|Scaridae)",
    fish.data)
```

and get all the species of gobies

```
goby <- familySearch("Gobiidae", fish.data)
```

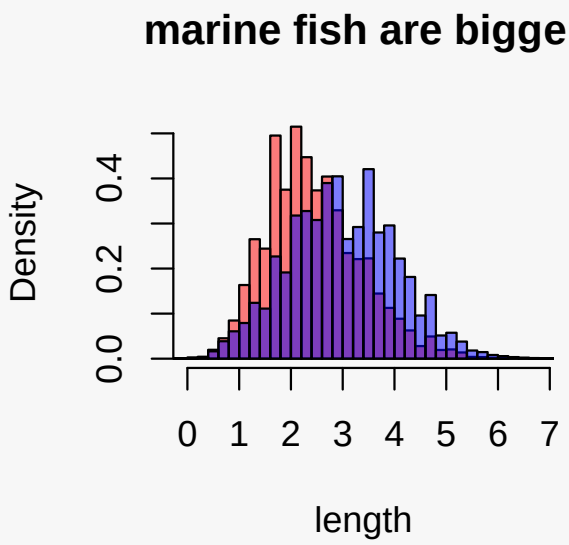
Identify how many labrids are found on reefs

```
labrid.reef <- habitatSearch("reef", fish.data[labrid])
nlabrids <- sum(labrid.reef)
```

and how many gobies are found on reefs:

*Corresponding author.
Email address: cboettig@ucdavis.edu (Peter Wainwright)

```
hist(log(getSize(fish.data[habitatSearch("freshwater",
    fish.data)], "length")), col = rgb(1, 0, 0, 0.5), breaks = 40,
    freq = F, xlab = "length", main = "marine fish are bigger")
hist(log(getSize(fish.data[habitatSearch("marine",
    fish.data)], "length")), col = rgb(0, 0, 1, 0.5), breaks = 40,
    add = T, freq = F)
```




```
ngobies <- sum(habitatSearch("reef", fish.data[goby]))
```

showing us that there are˜503 labrid species associated with reefs, and ˜401 goby species associated with reefs.

Note that any function can take a subset of the data, as specified by the square brackets.

2.2. Interfacing with other software

One of the greatest advantages about accessing fishbase directly through R is the ability to take advantage of the suite of specialized analyses available through R packages. Likewise, users familiar with these packages can more easily take advantage of the data available on fishbase. We illustrate this with an example that combines phylogenetic methods available in R with quantitative trait data available from `rfishbase`.

This series of commands illustrates testing for a phylogenetically corrected correlation between the maximum observed size of a species and the maximum observed depth at which it is found.

load a phylogenetic tree and some phylogenetics packages

```
data(labridtree)
require(geiger)
```

Find those species on FishBase

```
myfish <- findSpecies(tree$tip.label, fish.data)
```

Get the maxium depth of each species and sizes of each species:

```
depths <- getDepth(fish.data[myfish])[, "deep"]
size <- getSize(fish.data[myfish], "length")
```

Drop tips from the phylogeny for unmatched species.

```
data <- na.omit(data.frame(size, depths))
attach(treedata(tree, data))

Dropped tips from the tree because there were no matching names in the data:
 [1] "Anampses_geographicus"     "Bodianus_perditio"
 [3] "Chlorurus_bleekeri"        "Choerodon_cephalotes"
 [5] "Choerodon_venustus"        "Coris_batuensis"
 [7] "Diproctacanthus_xanthurus" "Halichoeres_melanurus"
 [9] "Halichoeres_miniatus"      "Halichoeres_nigrescens"
[11] "Macropharyngodon_choati"   "Oxycheilinus_digrammus"
[13] "Scarus_flavipectoralis"    "Scarus_rivulatus"

The following object(s) are masked _by_ '.GlobalEnv':

    data
```

Use phylogenetically independent contrasts˜[1] to determine if depth correlates with size after correcting for phylogeny:

```
x <- pic(data[["size"]], phy)
y <- pic(data[["depths"]], phy)
xtable::xtable(summary(lm(y ~ x - 1)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
x	0.0713	0.0993	0.72	0.4744

We can also estimate different evolutionary models for these traits to decide which best describes the data,

```
bm <- fitContinuous(phy, data[["depths"]], model = "BM")[[1]]
ou <- fitContinuous(phy, data[["depths"]], model = "OU")[[1]]
```

where the Brownian motion model has an AIC score of˜1185.3622 while the OU model has a score of˜918.158, suggesting that˜OU is the better model.

In a similar fashion, programmers of other R software packages can make use of the rfishbase package to make this data available to their functions, further increasing the use and impact of fishbase. For instance, the project OpenFisheries.org makes use of the fishbase package to provide information about commercially relevant species.

3. Discussion

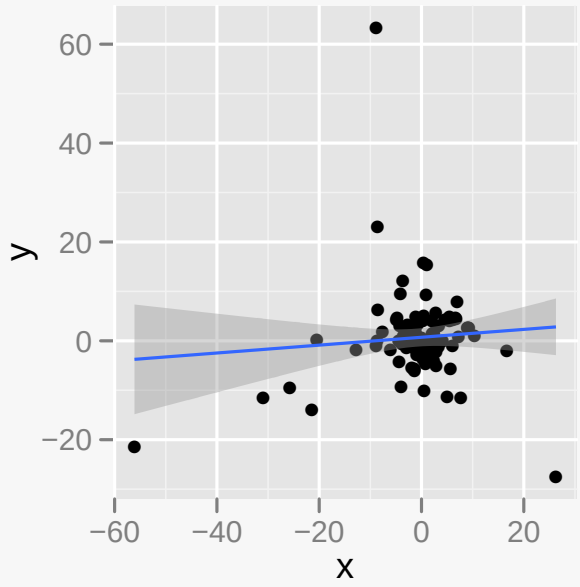
3.1. The self-updating study

Describe how this package could help make studies that could be automatically updated as the dataset is improved and expanded (like the examples in this document which are automatically run when the pdf is created). [5], [6].

3.2. Limitations and future directions

Fishbase contains much additional data that has not been made accessible in it's machine-readable XML format. We are in contact with the database managers and look forward to providing access to additional types of data as they become available.

```
ggplot(data.frame(x = x, y = y), aes(x, y)) +
    geom_point() + stat_smooth(method = lm)
```



4. Acknowledgements

CB is supported by a Computational Sciences Graduate Fellowship from the Department of Energy under grant number DE-FG02-97ER25308.

[1] Felsenstein, J., Jan. 1985. Phylogenies and the Comparative Method. The American Naturalist 125˜(1), 1–15.
URL http://www.journals.uchicago.edu/doi/abs/10.1086/284325

[2] Froese, R., Pauly., D., 2011. FishBase.
URL www.fishbase.org

[3] Hanson, B., Sugden, A., Alberts, B., Feb. 2011. Making data maximally available. Science (New York, N.Y.) 331˜(6018), 649.
URL http://www.ncbi.nlm.nih.gov/pubmed/21310971

[4] Jones, M.˜B., Schildhauer, M.˜P., Reichman, O., Bowers, S., Dec. 2006. The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere. Annual Review of Ecology, Evolution, and Systematics 37˜(1), 519–544.
URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.ecolsys.37.091305.110031

[5] Merali, Z., Oct. 2010. Computational science: Error. Nature 467˜(7317), 775–777.
URL http://www.nature.com/doifinder/10.1038/467775a

[6] Peng, R.˜D., Dec. 2011. Reproducible Research in Computational Science. Science 334˜(6060), 1226–1227.
URL http://www.sciencemag.org/cgi/doi/10.1126/science.1213847

[7] Reichman, O.˜J., Jones, M.˜B., Schildhauer, M.˜P., Feb. 2011. Challenges and Opportunities of Open Data in Ecology. Science 331˜(6018), 692–693.
URL http://www.sciencemag.org/cgi/doi/10.1126/science.1197962http://www.sciencemag.org/cgi/doi/10.1126/science.331.6018.692